Journal of Statistical Theory and Applications
Vol. **19(2)**, June 2020, *pp.* 286–296
DOI: https://doi.org/10.2991/jsta.d.200605.001; ISSN 1538-7887
https://www.atlantis-press.com/journals/jsta

On the Probabilistic Latent Semantic Analysis Generalization as the Singular Value Decomposition Probabilistic Image

Pau Figuera Vinué*, Pablo García Bringas

Faculty of Engineering, University of Deusto Unibertsitate Etorb., 24 Bilbo, Bizkaia 48007, Spain

ARTICLE INFO

Article History

Received 26 Mar 2019
Accepted 25 Apr 2020

Keywords

Singular value decomposition
Probabilistic latent semantic analysis
Nonnegative matrix factorization
Kullback–Leibler divergence

2010 Mathematics Subject Classification: 62H99, 15B48

ABSTRACT

The Probabilistic Latent Semantic Analysis has been related with the Singular Value Decomposition. Several problems occur when this comparative is done. Data class restrictions and the existence of several local optima mask the relation, being a formal analogy without any real significance. Moreover, the computational difficulty in terms of time and memory limits the technique applicability. In this work, we use the Nonnegative Matrix Factorization with the Kullback–Leibler divergence to prove, when the number of model components is enough and a limit condition is reached, that the Singular Value Decomposition and the Probabilistic Latent Semantic Analysis empirical distributions are arbitrary close. Under such conditions, the Nonnegative Matrix Factorization and the Probabilistic Latent Semantic Analysis equality is obtained. With this result, the Singular Value Decomposition of every nonnegative entries matrix converges to the general case Probabilistic Latent Semantic Analysis results and constitutes the unique probabilistic image. Moreover, a faster algorithm for the Probabilistic Latent Semantic Analysis is provided.

1. INTRODUCTION

The Probabilistic Latent Semantic Analysis (PLSA) have formal similarities with the Singular Values Decomposition (SVD). The obtained probabilistic factorization admits a matrix representation which can be assimilated to the SVD one. In fact, the original formulation of Hofmann [1] is a probabilistic remake of Latent Semantic Analysis (LSA), which is a direct SVD application introduced by Deerwester *et al.* [2]. Despite similarities, fundamental differences exist, since every real entries matrix admits the SVD decomposition, but only count data structures (and consequently contingency tables) PLSA decomposes. In fact, the PLSA has been derived from the probability theory, with hard restrictions. The idea of the PLSA is to factorize the co-occurrence table $N(w_i, d_j)$, identified as the relative frequencies $P(w_i, d_j)$ according multinomial distributions $P(w_i|\, z_k)$ and $P(d_j|\, z_k)$ with parameters θ_k and ϕ_k, being the $\mathbf{z}^t = (z_1, \ldots\, z_k)$ a set of categorical latent variables. Further, the model is adjusted until certain condition is achieved by the Expectation-Maximization (EM) algorithm. When the probabilistic formulas are written in the symmetric formulation, the formal analogy with the SVD is put in scene [1], but the assumptions restricts the PLSA to very special discrete cases, and limits seriously the conditions to establish the SVD-PLSA relation.

The work of Hofmann received strong criticism after publishing the main ideas on [1]. One of the most important one is the paper of Blei *et al.* [3], who pointed out several objections. The most important for the purposes of this manuscript is about the number of parameters of the PLSA model, which grows linearly when new data are added (documents), and the EM algorithm heritage, which has serious computational problems in terms of memory and execution time. Despite the problems, the PLSA is a very effective tool for information retrieval, text and image classification, bioinformatics, unsupervised machine learning, and constitutes a start point of some ideas as the probabilistic clustering. Despite the problems, the number of works using this technique has grown while the computing power increases.

A closely related technique to the PLSA is the Nonnegative Matrix Factorization (NMF). The NMF historically attributed to Chen [4] and extended by Paatero and Taper [5], can be used to solve similar problems to the PLSA in a less restrictive framework. The equivalence among the NMF and the PLSA has been discussed by Gaussier and Goutte [6] and Ding *et al.* [7], since every nonnegative entries matrix has a probabilistic interpretation under suitable normalization conditions, as is shown in [7], between many others. And advantage of the NMF is the ability to handle more general data structures than the discrete ones, referred by the PLSA. However, the PLSA solves the problem of maximizing the log-likelihood, while the NMF try to find two matrices $\mathbf{W}$ and $\mathbf{W}$ such that $\mathbf{X} \approx \mathbf{W}\,\mathbf{H}$ difference is minimal in some norm or divergence (a norm which does not satisfy all the distance axioms.)

*Corresponding author. Email: pau.figuera@opendeusto.es

While the PLSA practitioners are going on an algorithms competition to achieve good and reasonable fast results, not many theoretical works has been published. Some exceptions are the contributions of Brants [8] who uses the model for inferential purposes. Chaudhuri (2012) relates the diagonal matrix with the Mahalanobis distance [9], and Devarajan *et al.* [10] discuses several divergences, relating them with probability distributions.

The NMF create a scenario on which it is possible to derive similar equations to the obtained by Hofmann, for a wider data class, into a less restrictive framework. Deriving equations from the NMF with Kullback–Leibler (KL) divergence, ensuring the minimal number of components which guarantees monotonicity as is shown by Latecki *et al.* [11], and extracting the diagonal matrix as the inertia vector values, we call the formulas obtained in this way general case PLSA equations.

To avoid a bad definite problem or the undesirable effect of local maximums, the iterative procedure in the generalized PLSA factorization is done until the empirical distribution of the obtained product matrices is the same as the original data matrix, both with suitable normalization conditions. At this point, is easy to prove that for any nonnegative entries matrix, the singular value decomposition of the Gram–Schmidt orthonormalization (referred to it in the form given in [12, p. 24]) is the same as the generalized PLSA obtained matrices product Gram–Schmidt orthonormalization. Under those assumption is possible to establish the conditions which provides equal results, under the condition of the empirical distributions are reached. Then, the generalized PLSA formulas solves the PLSA (also, classical PLSA) problem too.

The next section is an overview of the main ideas, like the SVD theorem and the LSA, which constitutes the basis to introduce the PLSA as a probabilistic remake. The most fundamental ideas of the NMF and the PLSA formulas are introduced in Section 2 and Section 3 is devoted to derive the general case PLSA equations. Other section is devoted to the PLSA and SVD relation in the simplest way we can. Section 5 offers an example of the properties, with a reduced data set. Despite the example is driven with a small dimensions matrix, up to 10^5 iterations are needed to obtain acceptable results.

2. BACKGROUND

2.1. The Singular Value Decomposition

The well-known SVD is a generalization of the elementary lineal algebra Spectral Decomposition Theorem. The SVD states that every matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ can be decomposed as

$$\mathbf{X} = \mathbf{U}\,\Sigma\,\mathbf{V}^t \tag{1}$$

being $\mathbf{U}$ and $\mathbf{V}$ the orthogonal eigenvectors matrices of X and X^t, respectively. The $\sigma_1 \geq \cdots \geq \sigma_p \in \Sigma$ are the corresponding eigenvalues. If the σ_i values are sorted in decreasing order (to avoid the $p!$ columns permutations feasible results) the result is the same and the SVD decomposition is unique.

If the rank of $\mathbf{X}$ is p ($p < min(m, n)$), the approximate product $\hat{\mathbf{X}} \approx \mathbf{U}_r\,\Sigma_r\,\mathbf{V}_r^t \quad (r < p)$ is optimal in the sense of the Fröbenius norm which is given by $\| \mathbf{Y} - \hat{\mathbf{Y}} \|_F = \sum_r \sigma_r^2$. This is also called the dimension reduction problem.

The theoretical and practical consequences of the SVD are vaste. It is the main idea of many of the Multivariate Methods, and provides a geometrical interpretative basis for them. Moreover, the SVD does not involve many calculations and it is implemented in several of the existing programming languages [13, p. 309]. The flexibility and aptitude of the SVD for data analysis highlights the LSA, which can be considered a special interpretation case of the SVD.

On the LSA problem, the matrix $\mathbf{X}$ takes values on the positive integers and it is a count matrix. The elements of the diagonal matrix Σ are assimilated to a latent class of categorical hidden variables with any inferential sense [2]. A understandable latent class interpretation, based on text classification, could be when the rows of $\mathbf{X}$ are the words count over the documents placed in the columns. In this case, a reasonable interpretation of the latent class significance can be the analyzed documents subjects.

2.2. The PLSA Model

The PLSA model is a probabilistic remake of the LSA, obtained from the frequencies table $N(d_i, w_j)$. The main idea is to provide probabilistic sense to a count table of the words w_j in the documents d_i. By dividing each element of the table by the overall sum of the table, the relative frequencies $P(w_j, d_i)$ are obtained, and

$$P(w_j, d_i) = \frac{N(d_i, w_j)}{\sum_{ij} N(d_i, w_j)} \tag{2}$$

The expression Eq. (2) can be written as

$$P(w_j, d_i) = \sum_k P(w_j|\, z_k)\, P(z_k)\, P(z_k|\, d_i) \tag{3}$$

which is the asymmetric formulation of the PLSA [1][1] being z_k a set of K qualitative latent variables.

The PLSA estimates the probabilities of Eq. (3) maximizing the log-likelihood by using the EM algorithm. For the symmetric formulation is

$$\begin{aligned} \mathcal{L} &= \sum_{ij} \log \left(\mathbf{N}^{\, n(d_i, w_j)} \right) \\ &= \sum_{ij} n(d_i, w_j)\, \log \left(P(w_j|z_k)\, P(z_k)\, P(z_k|d_i) \right) \end{aligned} \tag{4}$$

Hofmann maximizes the expression Eq. (4) by computing the expectation, which is the posterior (E-step)

$$P(z|\, d, w) = \frac{P(z)\, P(d|\, z)\, P(w|\, z)}{P(z')\, P(d|\, z')\, P(w|\, z')} \tag{5}$$

and writes the Lagrangian, takes the derivatives, equalizes to zero, and eliminates the multipliers to maximize the expected log-likelihood (M-step). The provided solutions are [1]

$$P(w_j|z_k) = \frac{\sum_i n(d_i, w_j)\, P(z_k|w_j, d_i)}{\sum_{ij} n(d_i, w_j)\, P(z_k|w_j, d_i)} \tag{6}$$

$$P(z_k) = \frac{\sum_{ij} n(d_i, w_j)\, P(z_k|w_j, d_i)}{\sum_{ijk} n(d_i, w_j)\, P(z_k|w_j, d_i)} \tag{7}$$

$$P(d_i|z_k) = \frac{\sum_j n(d_i, w_j)\, P(z_k|w_j, d_i)}{\sum_{ij} n(d_i, w_j)\, P(z_k|w_j, d_i)} \tag{8}$$

By switching between Eq. (5) and the group of Eqs. (6–8), until certain condition is achieved, the model is adjusted.

The analogy between Eqs. (6–8) and the given by Eq. (1) is put in the scene when is written [1]

$$\mathbf{U} = [P(w_i|\, z_k)]_{ik} \tag{9}$$

$$\Sigma = \text{diag}\,(z_k) \tag{10}$$

$$\mathbf{V} = [P(d_j|\, z_k)]_{jk} \tag{11}$$

2.3. The NMF Point of View

A closely related PLSA problem is the MFN. It is usually stated as the decomposition of the nonnegative matrix $\mathbf{X} \in \mathfrak{R}_+^{m \times n}$ (*i*) in the r-dimensional cone Γ contained in the positive orthand, which is spanned by the columns of $\mathbf{W}$, $\mathbf{H} \subseteq \Gamma_r$ and (*ii*) the matrix $\mathbf{X}$ can be written as

$$\mathbf{X} \approx \mathbf{W}\,\mathbf{H} \tag{12}$$

being the column vectors of $\mathbf{X}$ convex combination of $\mathbf{H}$ [4].

The problem of the NMF for a matrix $\mathbf{X}$ is an optimization problem which Lee and Seung [14] solves by using the generalized KL divergence or I divergence.

$$D_I(\mathbf{X} \parallel \mathbf{W}\,\mathbf{H}) = \sum_{ij} [\mathbf{X}]_{ij} \log \frac{[\mathbf{X}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} - [\mathbf{X}]_{ij} + [\mathbf{W}\,\mathbf{H}]_{ij} \tag{13}$$

providing the results [14]

$$w_{ik} \leftarrow w_{ik} \frac{[\mathbf{Y}\,\mathbf{H}^t]_{ik}}{[\mathbf{W}\,\mathbf{H}\,\mathbf{H}^t]_{ik}} \tag{14}$$

[1]Also, it can be decomposed as $P(\ w_j, d_i) = P(d_i) \sum_k P(w_j|z_k)\, P(z_k|d_i)$, which is the asymmetric formulation.

$$h_{kj} \leftarrow h_{kj} \frac{[\mathbf{W}^t \mathbf{Y}]_{kj}}{[\mathbf{W}^t \mathbf{W}\,\mathbf{H}]_{kj}} \tag{15}$$

The equivalence between the PLSA and NMF problems needs a diagonal matrix. Gaussier and Goutte [6] shows that introducing diagonal matrices $\mathbf{S}$ and $\mathbf{D}$ on the factorization given in [14], the product Eq. (12) can be written as

$$\mathbf{W}\,\mathbf{H} = \mathbf{W}\,\left(\mathbf{S}\,\mathbf{S}^{-1}\right)\,\left(\mathbf{D}\,\mathbf{D}^{-1}\right)\,\mathbf{H} = (\mathbf{W}\,\mathbf{S})\,\left(\mathbf{S}^{-1}\,\mathbf{D}\right)\,\left(\mathbf{D}^{-1}\,\mathbf{H}\right) = \hat{\mathbf{W}}\,\mathbf{Z}\,\hat{\mathbf{H}} \tag{16}$$

being $\mathbf{Z}$ the diagonal matrix. This reveals that the PLSA solves the NMF. Ding deals with this problem, reaching same conclusions, but pointing out that the equivalent solutions, are not the same, debt to the algorithm differences [7].

3. GENERAL CASE EQUATIONS

Equivalent formulas to Eqs. (14) and (15) can be derived from the KL divergence. For $\mathbf{X}_+ \in \Re^{m\times n}$, the transformation $\mathbf{Y} = \mathbf{X}/n_X$ (with $n_X = \sum_{ij} \mathbf{X}$), valued on $\Re^{m\times n}_{[0,1]}$ is similar to the bi-variate distribution $P(d_i, w_j)$ in the PLSA model (also, classical PLSA formulation). Stating the NMF as a KL minimization divergence

$$\begin{aligned} D_{KL}(\mathbf{Y} \parallel \mathbf{W}\,\mathbf{H}) &= \sum_{ij} [\mathbf{Y}]_{ij} \log \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} \\ &= \sum_{ij} \left([\mathbf{Y}]_{ij} \log[\mathbf{Y}]_{ij} - [\mathbf{Y}]_{ij} \log [\mathbf{W}\,\mathbf{H}]_{ij}\right) \end{aligned} \tag{17}$$

the solutions, with the Karush–Kuhn–Tucker (KKT) conditions are

$$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} [\mathbf{H}]^t_{kj} \right) \tag{18}$$

$$[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left([\mathbf{W}]^t_{ik} \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} \right) \tag{19}$$

where $\odot$ is the Hadamard product.

Switching between Eqs. (18) and (19) after initializing, until certain condition is achieved, and assuming convergence in the iterative process, the obtained matrices product of $\mathbf{W}\,\mathbf{H}$ is $\hat{\mathbf{Y}}$ which estimates $\mathbf{Y}$ [2].

The pairs of Eqs. (18, 19), and (14, 15) are similar, but not equivalent since the derivation context is different [16, Chap. 3]. In the other hand, the classical KL divergence definition evidences the log-likelihood similitude with the second term of Eq. (17), while the first is a constant. This shows the equivalence between the KL divergence minimization and the EM algorithm. Moreover, the use of the KL divergence has good theoretical properties in a simple framework.

Formal similitude between the Eqs. (18) and (19) and the classical PLSA given by the formulas Eqs. (6–8), requires a diagonal matrix, which plays the role of Eq. (7).

Since the centered matrix of $\mathbf{Y}$ is

$$[\overline{\mathbf{Y}}]_{ij} = [\mathbf{Y}]_{ij} - 1\,\overline{\mathbf{y}}_j \qquad (j = 1, \ldots, n) \tag{20}$$

where 1 is the ones matrix, and $\overline{\mathbf{y}}_j$ are the column means vector. Choosing the diagonal matrix as

$$\text{diag}\,(\mathbf{t}) = \frac{\text{diag}\left([\overline{\mathbf{Y}}]^t_{ij}\,[\overline{\mathbf{Y}}]_{ij}\right)}{\text{trace}\left([\overline{\mathbf{Y}}]^t_{ij}\,[\overline{\mathbf{Y}}]_{ij}\right)} \tag{21}$$

[2] To prove convergence, the cost function is

$$F = \sum_{ij} [\mathbf{Y}]_{ij} \log\,[\mathbf{W}\,\mathbf{H}]_{ij}$$

and exist a G function accomplishing

$$\begin{aligned} G(h, h') &\geq F(h) \\ G(h, h) &= F(h) \end{aligned}$$

which leads to the sequence $h^{p+1} = \arg\min\, G(h, h')$. The derivatives provides a monotonically decreasing sequence. This procedure is similar to the detailed in [14, 15].

introducing vector notation for the NMF matrices

$$[\mathbf{W}\,\mathbf{H}]_{ij} = \langle \mathbf{w}_i, \mathbf{h}_j \rangle \qquad \left(\text{s.t. } \mathbf{w}_k,\ \mathbf{h}_k \in \Re^K\right)$$

Simple algebra manipulations

$$\begin{aligned}\langle \mathbf{w}_i, \mathbf{h}_j \rangle &= \left\langle \frac{w_{ik'}}{\sqrt{t_{k'}}}\sqrt{t_{k'}}\ , \frac{h_{k'j}}{\sqrt{t_{k'}}}\sqrt{t_{k'}} \right\rangle \\ &= \left[\frac{w_{ik'}}{\sqrt{t_{k'}}}\right] \operatorname{diag}(t_{k'}) \left[\frac{h_{k'j}}{\sqrt{t_{k'}}}\right]\end{aligned}$$

and identifying

$$[\mathbf{G}]_{ik} = \left[\frac{w_{ik'}}{\sqrt{t_{k'}}}\right]_{ik} \tag{22}$$

$$[\mathbf{F}]^t_{jk} = \left[\frac{h_{k'j}}{\sqrt{t_{k'}}}\right]_{jk} \tag{23}$$

the formal similitude with Eq. (1) for the NMF formulas derivation as the PLSA does, is obtained.

While the SVD dimension reduction problem is bounded by the nonzero eigenvalues, the number of model components is a controversial point within the PLSA, which inherits the EM algorithm problems. There are no restrictions on the number of latent variables (number of nonzero entries in $\mathbf{t}$). Nevertheless, after certain number of iterations, small values of the KL divergence rate occurs, and the new matrices differs not significantly from the previous ones. In this way, the problem is stated as the minimal number of components which allows similar divergences values to the obtained with a greater number of model components, after enough iterations. This idea, developed in [11], requires a closer look of the expression Eq. (5) for the multivariate case. Rewritten it in matrix notation, the array of matrices

$$[\hat{\mathbf{Y}}]_{ij} = \sum_k \alpha_k\, [\mathbf{W}\,\mathbf{H}]_{ij} \qquad \left(\text{s.t. } \sum_k \alpha_k = 1\right) \tag{24}$$

represents $P(z|\, w, d)$ of Eq. (6–8).

Each one of the $\alpha_k\ \mathbf{W}\,\mathbf{H}$ matrices is an element of the scalar tensor (or matrix array) in the space $\Re^{m\times n\times K}$, and the log-likelihood is written now as

$$\mathcal{L} = \sum_{ij}\sum_k \alpha_k\ [\mathbf{W}\,\mathbf{H}]_{ij} \tag{25}$$

$$= \sum_{ij}\left(\sum_k [\mathbf{W}\,\mathbf{H}]_{ijk}\right) \tag{26}$$

Only the largest values of α_k are significant terms in the sum Eq. (25), and is necessary that

$$\sum_k \alpha_k\, [\mathbf{W}\,\mathbf{H}]_{ijk} = [\mathbf{G}]_{ik} \operatorname{diag}(t_k)\ [\mathbf{F}]^t_{jk} \qquad (k = 1, \ldots, K) \tag{27}$$

which only can be achieved for a particular of the α_k values. Those values are the entries of $\mathbf{t}$. The minimal number of them which ensures this condition, minimizes the KL divergence too in the same way as more components does.

Taking into account that the empirical density of $\mathbf{Y}$ is $\mathbf{Y}^{(\text{emp})}$, and being

$$\begin{aligned}\beta &= \inf\left(y_{ij}\right) && \left(y_{ij} \in [\mathbf{Y}]_{ij}\right) \\ \beta^* &= \inf\left(y^*_{ij}\right) = \frac{1}{m\times n} && \left(y^*_{ij} \in [\mathbf{Y}]^{(\text{emp})}_{ij}\right)\end{aligned}$$

accomplishes $\beta \le \beta^*$. Since $\beta^* \to \beta$ while iterating, the inequality

$$\frac{K}{\min(m\times n)}\,\beta \ge \beta$$

holds only if

Proposition 3.1. *The number of model components which ensures that the empirical distribution of* $[\hat{\mathbf{Y}}]_{ij}$ *is the same as* $[\mathbf{Y}]_{ij}$ *is given by*

$$K \geq \min(m, n)$$

Introducing as a definition to refer the obtained formulas in a more concise way

Definition 3.1. For $\mathbf{Y}_+ \in \Re^{m\times n}$ such that $\sum_{ij}[\mathbf{Y}]_{ij} = 1$, the general case PLSA formulas are given by Eqs. (21–23). The number of model components is such that $K \geq \min(m, n)$. The entries of diag($\mathbf{t}$) are decreasing values ordered, and the same permutation is done in the columns of $\mathbf{G}$ and $\mathbf{F}$. In this case can be written

$$[\mathbf{W}\,\mathbf{H}]_{ij} = [\mathbf{G}]_{ik}\,\mathrm{diag}(\mathbf{t})\,[\mathbf{F}]^{t}_{jk}$$

and the sum $\sum_{ij}[\mathbf{W}\,\mathbf{H}]_{ij} = \sum_{ij}[\mathbf{G}]_{ik}\,\mathrm{diag}(\mathbf{t})\,\mathbf{F}^{t}_{jk} = 1$ is preserved.

4. PROPERTIES

The general case PLSA equations are a merely similitude reformulation of the Eqs. (21–23) with the SVD, if the connection between them is not established. The underlying relation can be seen by taking into account the convergence of the iterative process. Then, if $\mathbf{Y}^{(\mathrm{emp})}$ and $\hat{\mathbf{Y}}^{(\mathrm{emp})}$ are the empirical distributions of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ respectively, its accomplished that

Proposition 4.1. *The SVD of the orthonormalization (ortogonalization and column normalization) Gram–Schmidt of* $\mathbf{Y}$*, and the product of the general case PLSA equations* $\hat{\mathbf{Y}}$ *provides the same result when the condition* $\hat{\mathbf{Y}}^{(emp)} = \mathbf{Y}^{(emp)}$ *is reached.*

Proof. After orthogonalize and normalize $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, and denoting by ϕ the SVD decomposition, suppose that existes ϕ_1 and ϕ_2 such that

$$\phi_1(\mathbf{Y}) \sim \phi_1\left(\mathbf{Y}^{(\mathrm{emp})}\right)$$
$$\phi_2(\hat{\mathbf{Y}}) \sim \phi_2\left(\hat{\mathbf{Y}}^{(\mathrm{emp})}\right)$$

since $\mathbf{Y}^{(\mathrm{emp})} \sim \hat{\mathbf{Y}}^{(\mathrm{emp})}$, it follows $\phi_1(\mathbf{Y}) \sim \phi_2(\hat{\mathbf{Y}})$.

A direct consequence is, despite the data are rarely orthogonal

Proposition 4.2. *If the columns of* $\mathbf{Y}$ *are orthogonal, the SVD and the PLSA general case decompositions are the same.*

A practical way to ensure the empirical distribution equality $\hat{\mathbf{Y}}^{(\mathrm{emp})} = \mathbf{Y}^{(\mathrm{emp})}$ is reached, is to consider the Δ matrix operator. This operator compares the characters matrices $\mathbf{N}$ and $\hat{\mathbf{N}}$ obtained when a set of m labels substitutes the values of the columns, and they are arranged in decreasing order according the numerical values of the entries of the obtained matrix (it can be increasing too, with no differences on the results.) Those matrices can be seen as ordinal-named. Then, introducing as a definition

Definition 4.1. The Δ matrix operator is

$$\Delta_{ij}^{(p',p)} = \begin{cases} 0 & \text{if} \quad n_{ij}^{(p')} = n_{ij}^{(p)} \quad \left(p' > p \quad \text{and} \quad n_{ij}^{(p)} \in \mathbf{N}^{(p)}\right) \\ 1 & \text{otherwise} \end{cases} \tag{28}$$

being p and p' the iterations over which the comparison is done, and $n_{ij}^{(\cdot)} \in \mathbf{N}$ results from the column values substitution ordered according their entries.

The degree of adjustment between the characters matrices can be measured with the $\|\,\Delta_{ij}\,\|_1$ norm, which provides the number of nonco-incidences between them. When it is zero, the total adjustment between both empirical matrices is obtained, and a consequence is

Proposition 4.3. *The condition* $\|\,\Delta_{ij}\,\|_1 = 0$ *is equivalent to* $\hat{\mathbf{Y}}^{(emp)} = \mathbf{Y}^{(emp)}$.

In practical conditions the zero bound is difficult to achieve, since the matrix $\mathbf{Y}$ can contains some identical entries. In this case the labels ordination admits permutations on the repeated values, and the lower bound is the number of repeated values, named as r. Also, the row labels can be substituted by the column labels, with identical results. In the case when the lower bound of $\Delta = r$ is achieved, it can be stated that

Proposition 4.4. *The matrix* $\hat{\mathbf{Y}}$ *reconstructs* $\mathbf{Y}$ *(and* $\mathbf{X}$ *if the total sum has been saved).*

Proof. If $\hat{\mathbf{Y}}^{(\mathrm{emp})} = \mathbf{Y}^{(\mathrm{emp})}$ and denoting by $\widetilde{\mathbf{Y}}^{(\mathrm{emp})}$ the column normalized matrix of $\mathbf{Y}^{(\mathrm{emp})}$

$$\widetilde{\hat{\mathbf{Y}}}^{(\mathrm{emp})} = \widetilde{\mathbf{Y}}$$

and after dividing for the total row and column sum again for both matrices, with an round-off approximation error $\epsilon \leq (m \times n)$, the equality is proved.

In this case the maximum is achieved, since from a numerical point of view, when the condition $\Delta^{(p,0)}$ is reached, the approximation error between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ is small. In this case the surfaces defined by the two matrices are similar, and they reproduces all the extremes vales and modes.

A similar construction to the given by propositions 4.1–4.4 can be build up for the classical PLSA Eqs. (6–8), when they are written in matrix form, reaching also the maximum. In this case the relation between the classical PLSA model and the general case PLSA equations relies on the significance of the diagonal matrix given by Eqs. (6) and (21), respectively. In this case the relation between $P(z)$ and $\mathbf{t}$ is, and

Proposition 4.5. *The expectation of* $P(\mathbf{z})$ *is*

$$E[P(\mathbf{z})] = \frac{diag\ (var(\mathbf{Y}))}{trace\ (var(\mathbf{Y}))}$$

Proof. Since $\mathbf{z} = P(z_k) \quad (k = 1, \ldots, K)$

$$\mathrm{E}[P(\mathbf{z})] = \sum_k P(\mathbf{z})\, z_k = \mathbf{z}^t\, \mathbf{z}$$

and expressing as in Eq. (24), since the α's simplifies

$$\mathbf{z}^t\, \mathbf{z} = \frac{\sum_{i\,k'} [\mathbf{W}\,\mathbf{H}]^t \sum_{k'j} [\mathbf{W}\,\mathbf{H}]}{\sum_{ij\,k'} [\mathbf{W}\,\mathbf{H}]^t \sum_{ijk'} [\mathbf{W}\,\mathbf{H}]} = \frac{\mathrm{diag}\left(\mathbf{Y}^t\,\mathbf{Y}\right)}{\mathrm{trace}\left(\mathbf{Y}^t\,\mathbf{Y}\right)} = \frac{\mathrm{diag}\ (\mathrm{var}\,(\mathbf{Y}))}{\mathrm{trace}\ (\mathrm{var}\,(\mathbf{Y}))}$$

Introducing as a definition the steps to ensure the equality between decompositions

Definition 4.2. The probabilistic SVD image is obtained when the general case PLSA equations reach the limit on which the empirical distributions are the same, except for r repeated values of the data matrix $\mathbf{Y}_{[0,1]}$ obtained from $\mathbf{X}_+$

And recompiling the previous properties as a compact result, it can be stated

Theorem 4.1. *The probabilistic SVD image, the classical and general case PLSA matrix factorization are equal when* $\Delta = r$ *(being r the number of repeated values in* $\mathbf{Y}$*). The orthonormalized SVD is the same for them. In this case the local basis which spans the general case PLSA equations is the orthonormalized basis which spans too the transformed data matrix* $\mathbf{Y}$*, obtained from* $\mathbf{X}$ *by dividing for the overall sum.*

If statistical inference will be done is unique necessary to normalize suitable columns or rows. In this work we go no far on this point.

5. EXAMPLE

This section offers a very simple example. It analyzes the effect of the number of components in the general case PLSA equations on the reached convergence limit, and compares it with the classical PLSA. For this purpose the Δ matrix goodness of use is examined. Both models offers similar numerical results.

Since the number of iterations which are necessary to reach the maximum grows linearly with the data matrix dimension and the number of components, the small data set decathlon is used to drive the example. Included in several R packages, with some differences among them, the selected one is the included in the *FactoMineR* one [17]. The data are the ranks of elite athletes participants in the Athenas 2012 Olimpic Games men's decathlon competition. Additional reductions are done in the data by selecting 28 rows and 10 columns from the 41 and 13 original ones. Those are the athletes unique results only in the Olympic Games. Other meetings reference values are deprecated, and omitting total points and classification too. Thus, every row corresponds to a unique participant. The reduced data set is column labeled as 100 (100 meters), long (long jump), poid (shot put), haut (high jump), 400 (400 meters), 110 (110-meter hurdles), disc (discus throw), perc (pole vault), jave (javelin), and 1500 (1,500 meters). The athletes name are the rows labels, used as identifiers.

The data can be written as a nonnegative real valued 28×10 matrix, denoted as $\mathbf{X}$. Every item (athlete) is a row vector. The column vectors are the total marks of the trials, and they are continuous variables. If the row and column names are preserved, the transformation $\mathbf{X}/n \left(n = \sum_{ij}\right) \mathbf{X}$ provides the matrix $\mathbf{Y}$. The correspondent qualitative or chars matrix $\mathbf{N}_c$ corresponding to $\mathbf{Y}$ is obtained by substituting in every column the athlete name instead of the mark, and reordering into it according the obtained mark (which is the trial classification).

One must be careful at the moment to ordering the results, if significance will be provided. There are two categories: more is better, which corresponds to the distance achieved for events like jumping or shots, and less is better, which is the case of the time to cover a distance by

running. Not always this correspondence occurs, and it should be done in a cautious way in all the cases if significance will be provided, but it is not important for computation purposes. The qualitative matrix $\mathbf{N}_c$ has not algebraic sense. The ordination according to the obtained ranks can be omitted and a ascendant or descendant one is sufficient for comparison purposes. This task is left to the analyst criteria and has not more importance than coherence with the data ordination.

To see how the Δ matrix works, running the equations of Definition 3.1 from initial random conditions, 10 model components, and $p = 1,000$ iterations, a estimation $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ defines $\mathbf{N}_c^{(1000)}$. This qualitative matrix is obtained by substituting the numbers of the rows by the row-label (or athlete name). Something similar to the Table 1 will be obtained.

The comparation between $\mathbf{N}_c$ and $\hat{\mathbf{N}}_c^{(1000)}$ is done according to the criteria given by Definition 4.1, the Δ matrix for the results of Table 1 is

$$\Delta_{ij}^{(1000,\,0)} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}$$

being the zeros the coincidences of the rank of the athlete in the column trial and the ones appears when they are different. The L_1 norm gives the accuracy of the classification. In an ideal case it should be zero, but repeated values appear. In this case, the 47 repeated values in the data matrix $\mathbf{Y}$ provides $r = 47$ permutations of the character matrix, being indistinguishable between them, and this is a limit for Δ.

The objective is to adjust the Δ matrix to $r = 47$ noncoincidences (or less). Then, the minimal number of model components to ensure this condition K, given by the Proposition 3.1. For realistic data sets, this condition is difficult to achieve, since is expensive in terms of computational resources, since the number of iterations is not small, as it can be seen in Figure 1.

From a purely computational point of view, the obtained results are compared with the classical PLSA, which offers similar results as those shown in the Figure 1. In this case the PLSA is improperly run, since the data are not multinomial distributed, but the purpose is purely computational. The results are the same, and differences are debt to random initialization conditions. In the Figure 1 it can be seen too how oscillates close to the maximum, and a little more of iterations stabilizes the result. At this point the numerical approximation error is small and lets to reconstruct the original empirical distribution, and if the sums are saved, the original data matrix can be reconstructed (except for the repeated values and labels of the data matrix).

For a enough large number of iterations, it can be seen how the diagonal matrix obtained by the formula (21), which is the estimation of $P(z_k)$ defined in Eq.(6), is the expectation of $P(z)$, as is shown in the Figure 2. To obtain the graphics, the classical PLSA algorithm has been executed 20 times for a uniform distribution and 20 times for a Poisson initial distributions of the latent variables z_k. Its clear the effect of the initial conditions on the obtained results. Being they equivalent, are not the same. The solution proposed by Eq. (21) is a limit distribution in each case, and ensures a well-based probabilistic sense for the diagonal matrix.

A not minor consequence of the general case PLSA formulas is the increasing computational speed, as can be seen in Table 2. A well-known problem of the PLSA practitioners is the complexity in terms of time and memory. The general case PLSA equations reduces this drawback to a NMF difficulty one. Since the EM algorithm convergence is slow, the KL based too, and it is a consequence of their intimately connection, but the fact to avoid the estimation of the object of Eq. (5), simplifies the operations. It has important consequences, since the final results requires less time. When only in PLSA practical results interest is, a trick is to run Eqs. (18) and (19) with a large number of model components, since the interest relies only in matrices $\mathbf{W}$ and $\mathbf{H}$. When those matrices are obtained, it is easy to get the diagonal by using Eq. (21) with the desired number of model components. This procedure is hard to justify, but in practice works well.

Table 1 Obtained qualitative matrices of characters.

	100	**Long**	**…**	**1500**
	LB	**MB**	**…**	**LB**
$\mathbf{N}_c$	Karpov	Clay	…	Lorenzo
	Averyanov	Sebrle	…	Smirnov
	Warners	Karpov	…	Hernu
	⋮	⋮	⋮	⋮
	Schoenbeck	Parkhomenko	…	Korkizoglou
$\mathbf{N}_c^{(1000)}$	Karpov	Clay	…	Lorenzo
	Clay	Sebrle	…	Gomez
	Averyanov	Karpov	…	Smirnov
	⋮	⋮	⋮	⋮
	Casarsa	Casarsa	…	Korkizoglou

Note: Characters matrix $\mathbf{N}_c$ corresponding to the data matrix $\mathbf{Y}$ and characters matrix $\mathbf{N}_c^{(1000)}$ of $\hat{\mathbf{Y}}$ after $p = 1,000$ iterations.

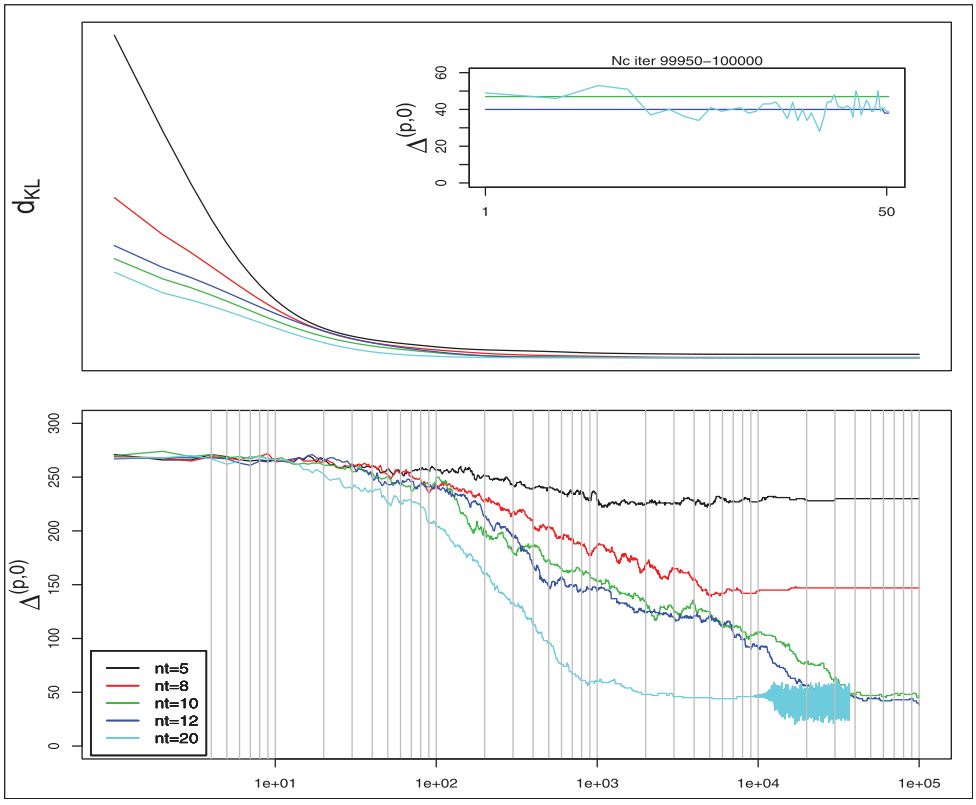

Figure 1 The two figures share the same abscissa axis. The top graphics shows the how decreases the KL divergence as the iterations increases for different number of model components. Initial values are randomized. The bottom figure shows the number of mismatches. When Δ matrix has 47 non-coincidences, oscillates around this value, revealing that is close to the maximum. The window of the top figure shows this fact. nt is the number of model components.

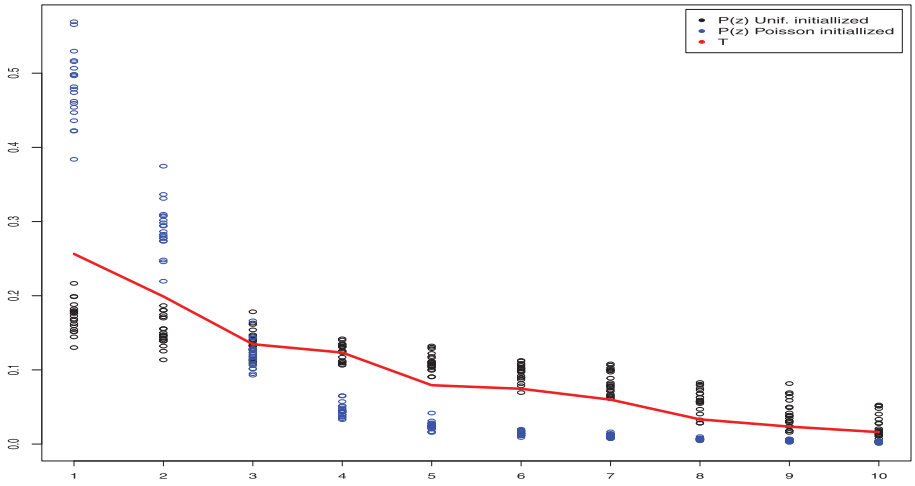

Figure 2 Expected value for $P(\mathbf{z})$ when the classical PLSA algorithm has been initialized under different conditions. The uniform initialized $P(\mathbf{z})$ values has the limit exponential distribution [18, p.157] and the Poisson, under certain conditions is a exponential too [18, p.158]. Both cases are a special selection parameters of a χ^2 distribution [19]. T represents diag (t).

6. DISCUSSION

The iterative process to obtain the generalized PLSA matrices, minimizes the KL divergence between the object and image matrix. The entries of $\mathbf{G}$ diag $\left(\mathbf{t}^{1/2}\right)$ are spanned over the basis diag $\left(\mathbf{t}^{1/2}\right)\mathbf{F}$ (or vice versa, depending on the desired interpretation). In the optimal neighborhood, this local basis representation has the same coordinates as the classical algebraic decomposition. Taking into account this

Table 2 Speed comparison between classical and generalized Probabilistic Latent Semantic Analysis (PLSA) equations algorithm performance (time for 10^4 iterations).

	Time Cicle (ms)	Total Memory (MB)
Classical PLSA	14,379	132.0
Generalized PLSA	1,350	106.6

Note: Computations done with a Intel 8.00 GB RAM 2.30 GHz processor.

fact, the projective distances in the L_2 space norm are also minimized. This consequence should be analyzed in more depth, to reveal the relation between algebraic structures and contained information, since KL divergence is a measure of information.

Another point to study in more depth are the local basis rotations from a statistical point of view, to ensure the results interchangeability between both points of view. In the euclidean space case, this fact is depth studied, and the consequences are irrelevant, but it is not so clear when talking about probability distributions.

The general case PLSA and the SVD relation has been established for the full rank case. The low rank is used in connection with the problem of dimension reduction. These concepts should be also related for the general case PLSA, to establish the limits and advantages of this similarity, and reveal in a more clear way the achieved convergence limit.

From a purely practical point of view, it is necessary to increase the computational and convergence speeds of the algorithm. The faster iterative procedure with the I-divergence is the basis of many of the current algorithms, but cannot be put in relation with the KL divergence. Necessarily, this leads to establishing the relationship between the KL divergence (or the EM algorithm) with other types of divergences. Therefore, is necessary a broader conceptual base on this field.

7. CONCLUSIONS

From the obtained results, three conclusions can be established. First, the probabilistic SVD image is asymptotically unique and allows to provide inferential sense to descriptive statistical techniques based on the SVD. Second, an algebraic consequence is the total equivalence between the classical PLSA and the general PLSA model (when certain conditions are satisfied). This shows that statistical problems, under suitable conditions, is an algebraic one, under a reliable conceptual basis. Finally, the general case PLSA equations has no distributional hypothesis, and solely restricted to real-valued entries. This leads to estimate under a quantitative basis the optimal number of model components and its significance.

CONFLICT OF INTEREST

No conflict of interest declaration for manuscript title.

AUTHORS' CONTRIBUTIONS

Figuera Pau: Conceptual development and main work. García Bringas Pablo: Professor Garcia Bringas is my Thesis Director, and his inestimable asportation is on fundamental questions and objections, with infinitely patience, to make clear the main ideas.

ACKNOWLEDGMENTS

We would like to show our gratitude to the editor-in-chief Prof. M. Ahsanullah and the unknown referees, for the patience to accept relevant changes once the manuscript was submitted. Without this collaboration, this manuscript would have been impossible to be published.

REFERENCES

1. T. Hofmann, J. Mach. Learn. Res. 42 (2000), 177–196.
2. S. Deerwester, S. Dumais, G. Furnas, *et al.*, J. Assoc. Inf. Sci. Technol. 41 (1990), 391–407.
3. D. Blei, A. Ng, M. Jordan, J. Mach. Learn. Res. 3 (2003), 993–1022.
4. J. Chen, Linear Algebra Appl. 62 (1984), 207–217.
5. M. Paatero, U. Taper, Environemetrics. 5 (1994), 111–126.

6. E. Gaussier, C. Goutte, in: R.A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, J. Tait (Eds.), SIGIR '05 Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2005, pp. 601–602.
7. C. Ding, T. Li, W. Peng, Comput. Stat. Data Anal. 52 (2008), 3913–3927.
8. T. Brants, Inf. Retrieval. 8 (2005), 181–196.
9. A. Chaudhuri, M. Murty, in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, Tsukuba, Japan, 2012, pp. 2298–2301.
10. K. Devarajan, G. Wang, N. Ebrahimi, Mach. Learn. 99 (2015), 137–163.
11. L. Latecki, M. Sobel, R. Lakaemper, in KDD'06 Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Minings, ACM, Philadelphia, PA, USA, 2006, pp. 267–276.
12. A. Khuri, Advanced Calculus with Applications in Statistics, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New Jersey, USA, 2003.
13. H. Golub, C. Van Loan, Matrix Computations, The Johns Hopkins University Press, Maryland, USA, 1996.
14. D. Lee, H. Seung, in: T. Leen, T. Dietterich, V. Tresp (Eds.), in 14th Annual Neural Information Processing Systems Conference (NIPS), MIT Press, Denver, CO, USA, 2000, pp. 556–562.
15. A. Dempster, N. Laird, D. Rubin, J. R. Stat. Soc. 39 (1977), 1–22.
16. A. Cichocki, R. Zdunek, A. Phan, S.-I. Amary, Nonnegative Matrix and Tensor Factorizations, John Willey and Sons, Ltd, Washington, USA, 2009.
17. S. Lê, J. Josse, F. Husson, J. Stat. Softw. 25 (2008), 1–18.
18. N. Balakrishnan, V. Nevzorov, A Primer on Statistical Distributions, John Wiley and Sons, 2005.
19. G. Casella, R.B.S. Edition, Statistical Inference, Duxbury, Massachusetts, USA, 2002.